CHAPTER

15 Observational Methods

Jamie M. Ostrov *and* Emily J. Hart

Abstract

Systematic observational methods require clearly defined codes, structured sampling and recording procedures, and are subject to rigorous psychometric analysis. We review best practices in each of these areas with attention to the application of these methods for addressing empirical questions that quantitative researchers may posit. Special focus is placed on the selection of appropriate observational methods and coding systems as well as on the analysis of reliability and validity. The use of technology to facilitate the collection and analysis of observational data is discussed. Ethical considerations and future directions are raised.

Key Words: Observation, observer, time sampling, event sampling, participant observation, focal participant sampling, semi-structured observations, scan sampling, interobserver reliability, Cohen's Kappa, observer drift, reactivity, remote audio-visual recording, computer-assisted observational software

Introduction

Systematic observational methods have been a common technique employed by psychologists studying human and animal behavior since the inception of our field, and yet best practices for the use of observational instruments (*see* Table 15.1) are often not known or adopted by researchers in our field. As such, the quality of observational research varies widely, and thus, it is our goal in the present chapter to review and explicitly define the standards of practice for this important methodological tool in the psychological sciences. Bakeman and Gottman (1987) have previously defined observational methods to include the *a priori* use of operationally defined behavioral codes by observers who have achieved interobserver reliability. Importantly, the setting or context is not what defines a method as being systematic (Pellegrini, 2004). That is, systematic observations may be conducted in the laboratory, schools, workplace, public spaces and coded live or via recordings/transcripts. Therefore, having clear definitions and sampling/recording rules as well as reliable codes delineates informal, unsystematic observation from systematic observation. We also distinguish between the use of nonsystematic field notes and other data collection techniques that are often used in qualitative studies by ethologists and educational practitioners in naturalistic contexts and only include a review and analysis of systematic observational methods (Pellegrini, Ostrov, Roseth, Solberg, & Dupuis, in press).

Nonsystematic sampling techniques such as *Ad libitum* (i.e., ad lib) in which there are no *a priori* systematic sampling or recording rules are often used by researchers as a part of pilot testing and help to inform the development of systematic observational coding systems (Pellegrini, 2004). Thus, ad lib sampling approaches are important to understand the context and nature of the behaviors under study, but they will not be discussed

Table 15.1. Best Practices for Observational Methods

Methodological issue	Best practice recommendation
Defining behaviors/ codes	Clear, discrete behaviors are ideal. *A priori* operational definitions/observational codes are needed. Codes should be mutually exclusive and exhaustive where appropriate.
Sampling/ recording rules	Procedures should be standardized and appropriate for the behavior under study. Observations should be independent and pilot-tested if a new scheme is used.
Training	Observers should be unaware of study hypotheses. A standardized manual should be used. Initial levels of interobserver reliability should be obtained by all observers with an experienced, reliable trainer.
Data collection	A minimally responsive manner should be used for live observations to reduce participant reactivity. Participants are only observed once per session/day.
Reliability/ validity	Interobserver reliability should be assessed across the study. Cohen's Kappa should be used when possible. Validity assessments should be included.
Scoring	Standardized procedures should be adopted.
Biases/Error	Efforts should be implemented to reduce participant reactivity, observer drift, and other biases and sources of error.
Ethics	IRB approval as well as informed consent/assent should be obtained when possible. Protections should be considered for the duration of the study.

further in this review. Observational methods may be used in a variety of designs from correlational and quasi-experimental to experimental and even randomized trial designs (Bakeman & Gnisci, 2006). However, it is more typical to find systematic observational methods used outside the laboratory to maximize ecological validity and, thus, less likely as part of experimental manipulations (Bakeman & Gnisci, 2006). The current review will be relevant to all research designs with a focus on those methods that are well designed for quantitative data analysis.

History of Observational Methods

The use of systematic observational methods has been used extensively by psychologists throughout the history of our field to examine various empirical questions (*see* Langfeld, 1913). One of the first documented cases of systematic observational methods in the extant literature was from a study by Goodenough (1930) and was part of an increasing trend in the systematic study of young children as part of the Child Welfare Movement in the United States, which was supported by the National Research Council (for review, *see* Arrington, 1943). In fact, her seminal work was also one of the first studies in psychology to be published using time sampling (*see Sampling* section below) observational procedures (Arrington, 1943). In her classic work (appearing in the first issue of *Child Development*), Florence L. Goodenough reported on several observational studies conducted in her laboratory at the Institute of Child Welfare (now Institute of Child Development) of the University of Minnesota. This study highlights several best practices that are still endorsed today. For example, careful pilot testing of the observational codes was conducted, and revisions were made to generate mutually exclusive codes (*see Coding* section below) and reliable distinctions between the categories. In addition, observations of each child's physical activity were conducted only once per day and only by one observer at a time so that observations of behavior were conducted independent of one another. Goodenough (1930) carefully defined the *a priori* categories or observational codes and demonstrated interobserver reliability for each of these codes. Finally, Goodenough (1930) described the justification for her observational procedures and discussed alternative techniques (e.g., the optimum duration for an interval within a time-sampling procedure). There are other well-known examples of systematic observation conducted by contemporaries of

Goodenough, including Parten's (1932) study of young children's play behavior, which also illustrate best practices (e.g., clearly defined, mutually exclusive observational codes; rules designed to maintain independence of sampling and decrease observer error). Some of the earliest observational studies focused on either children or non-human animals (e.g., Crawford, 1942), as other techniques for studying behavior (and often social domains of study) were either not as well suited for the research questions or not available at the time. Today, systematic observational methods are used in research and applied settings (Pellegrini, 2001) and relevant for training in all domains and subdisciplines of the social and behavioral sciences (Krehbiel & Lewis, 1994).

Sampling and Recording Rules

Systematic observational systems follow various sampling and recording rules that are designed for different contexts and research questions. The following section includes a review of the central sampling and recording rules that quantitative scholars would use for conducting systematic observations (*see* Table 15.2 for a summary of the strengths and weaknesses of each approach). Recently adopted best practices for direct systematic observation are relevant for each of these types of observational methods, and they are briefly reviewed here. These practices, which were first introduced by Hintze, Volpe, and Shapiro (2002), include (1) the observational system is designed to measure well-defined behaviors; (2) the behaviors are operationally defined *a priori*; (3) observations are recorded using objective, standardized (i.e., manualized training protocols) sampling procedures and recording rules; (4) the context and timing of sampling is explicitly determined; and (5) scoring and coding of data are conducted in a standardized fashion (*see* Leff & Lakin, 2005, p. 476).

Time Sampling

A time-dependent observational procedure in which the researcher *a priori* divides the behavior stream into discrete intervals and each time interval is scored for the presence or absence of the behavior in question is defined as a time sampling observational approach. That is, the time interval is the unit coded (Bakeman & Gottman, 1987). Time sampling procedures may be conceptualized as either 0/1 (i.e., absent/present or nonoccurrence/occurrence) or continuous in nature. A time sampling procedure is an efficient method of sampling, as multiple data points may be collected from a single participant in a short period of time. Time sampling is well suited for measuring rather discrete behaviors, such as overt behaviors (e.g., on task and off task behavior in classrooms), or with behaviors that are frequently occurring. For example, a recent study of the frequency of various behaviors (e.g., off task behavior, noncompliance) during several naturalistic activities in 30 children with various psychiatric diagnoses used a reliable 0/1 time sampling approach with a 15-second interval (Quake-Rapp, Miller, Ananthan, & Chiu, 2008). Alternatively, time sampling is not well designed for infrequently occurring events or events that are long in duration (Slee, 1987). A clear advantage is that time sampling is relatively inexpensive because it is an efficient use of the research assistant (Bakeman & Gottman, 1987). Further, 0/1 sampling is also easier for the observer than alternatives such as instantaneous sampling, in which the research assistant notes if the behavior is present at a precise moment in time rather than it occurring during a larger interval of time. A major disadvantage of the time sampling approach is that the researcher delineates the particular time interval and therefore arbitrarily categorizes the behavior into discrete artificial units of time that may or may not be meaningful (Slee, 1987). Moreover, some behaviors may exceed the often brief interval of time that is selected for the sampling. Thus, it is crucial to carefully justify the interval that is selected. The intervals are often brief and the behaviors in question should be readily apparent and easily observable by trained research assistants. If frequency estimates are to be obtained, then the interval in question needs to be sufficiently brief so that an accurate assessment can be made. That is, typically with an interval approach, a maximum of one behavior is recorded during an interval even if the behavior independently occurs more frequently during this interval (Slee, 1987). Thus, special attention needs to be given to the pilot testing of the observational scheme and various durations of the interval if frequency assessments are desired.

Time sampling procedures are used in a range of settings and studies to test various empirical questions that often have applied significance. For example, Macintosh and Dissanayake (2006) adopted a 0/1 time sampling technique to assess spontaneous social interactions in school-aged children with high-functioning autism or Asperger's disorder as well as typically developing children. Observations were conducted in the schoolyard. For each

Table 15.2. Strengths and Weaknesses of each Observational Approach

Method	Strengths	Weaknesses
Time sampling	This method is efficient and inexpensive. It is appropriate for frequently occurring and/or discrete behaviors.	It is less useful for infrequently occurring behaviors. Time units may be categorized inappropriately.
Event sampling	This method efficiently enables the measurement of frequency, duration, latency, and intensity. It may be used with frequently or infrequently occurring behaviors.	It may be inappropriate in situations where it is difficult to determine the independence of events, such as dyadic interactions.
Participant observation	This method is appropriate for the study of broad and complex constructs that encompass a variety of events or behaviors. It may be useful in applied settings.	It is less efficient.
Focal sampling	This method allows for in-depth recording of an individual participant. Continuous recording enables multiple types, sequences, and true frequencies of behaviors. May be useful in applied research contexts.	Large amounts of time are often needed.
Scan sampling	Instantaneous recording rules promote efficiency. It is appropriate for overt, readily observable behaviors.	It may be difficult to obtain true frequency of a behavior. It is less appropriate for subtle behaviors.
Semi-structured observations	Experimental control is provided.	Ecological validity may be lacking. It requires additional work to pilot test and validate the paradigm.

timed interval of 30 seconds, one type of behavior (e.g., parallel play) from a particular behavioral domain (e.g., social participation) was coded. For reliability purposes, a second observer made independent ratings for 20% of the entire sample. Intraclass correlation reliability coefficients were all acceptable for each type of behavior (0.78–0.99) with the exception of nonverbal interaction (i.e., gestures; 0.58), which are often difficult to reliably assess in live settings (*see also* Ostrov & Keating, 2004). Results meaningfully distinguished between the typically developing children and the clinical groups and revealed few differences between the two clinical groups, supporting the use of time sampling as a means to discriminate between clinical and nonclinical groups (Macintosh & Dissanayake, 2006). Time sampling procedures have several other applications and clinical considerations. For example, time sampling methods may differentially affect how treatment effects are interpreted (Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007) and may be appropriate for classroom-based research that tests adherence to educational policies intended to aid students with special needs (Jackson & Neel, 2006; Soukup, Wehmeyer, Bashinski, & Boyaird, 2007).

Event Sampling

Event-based sampling is also known as behavior sampling and permits a researcher to study the frequency, duration, latency, and intensity of the behavior under study (Pellegrini, 2004). Essentially, unlike time sampling, event sampling is a type of observational sampling in which the events are time-independent and the behavior is the unit of analysis (Bakeman & Gottman, 1987). Event sampling allows the behavior to remain as part of the naturally occurring phenomenon and may unfold in a manner generally consistent with the timing of the behavior in the natural setting. This type of sampling also can be efficient in terms of the total amount of time needed for observations. Unlike other sampling techniques (e.g., time sampling), a third advantage is that event sampling may be used when the construct under study is either frequently or infrequently occurring (Slee, 1987). There are some clear disadvantages to event-based sampling procedures,

and this may be a reason that it is less commonly seen in the literature. First, it is sometimes challenging to delineate the independence of events—that is, the researcher must specify when one event ends and the next event begins. Second, event sampling does not lend itself well to coding of dyadic interactions such as parent–child or romantic partner relations in which there is a fair amount of interdependence between the participants (Slee, 1987).

Event sampling also has wide applicability and has even been used to understand the propensity to violence at sporting events. For example, Bowker et al. (2009) used an event-sampling approach to examine spectator comments at youth hockey games in a large Canadian city. A group of five observers attended 69 hockey games played by youth in two age groups: 11–12 years and 13–14 years. Verbal comments were coded as positive, negative, corrective, or neutral and rated for intensity. Most of the comments elicited by spectators were positively toned. The valence of spectator comments was influenced by gender (i.e., the gender of the children playing) and the purpose for which the game was being played (i.e., competitive or recreational). These results support the utility of event sampling at social and athletic events, where particular behaviors are likely to occur during a finite period of time. Time sampling may not be appropriate in such circumstances because of the presence of a high concentration of individuals in a single setting and many potential interruptions arising from the nature of the activity.

Participant Observation

Although participant observation has been more frequently used with nonsystematic field observation and in disciplines that focus on qualitative methods, it is possible to conduct systematic participant observation as part of quantitative studies. Systematic participant observation has been the method of choice for behaviors of interest that require "an insider's perspective" (Pellegrini, 2004, p. 288) or for contexts in which the sampling period may be long and informal. Moreover, this method is well suited for the use of more global observational ratings that sample events. This procedure has wide applicability, and participant observation has an extensive history of successful use from studies of children with behavioral problems at summer camps in clinical psychology (e.g., Newcomb, 1931; Pelham et al., 2000) to worker stress in organizational psychology (e.g., Länsisalmi, Peiró, & Kivimäki, 2000). For example, a recent study of children diagnosed with disruptive behavioral disorders and enrolled in a summer treatment program used staff counselors to complete daily participant observations of social behaviors of the children while they engaged in various camp activities (Lopez-Williams et al., 2005). A second study of social competence among reunited adolescents (M age = 15.5 years) who had attended a research-based summer camp when they were 10 years old revealed the predictive validity of participant observer (i.e., camp counselor) ratings of social skills (Englund, Levy, Hyson, & Sroufe, 2000). The validity of the participant observations of social competence when the participants were 10 years old was determined by revealing significant prospective correlations with a group-problem solving task that was videotaped and coded by two independent raters along several dimensions (e.g., self-confidence, agency, overall social competence) when the participants were 15 years old. The results support the use of participant observations in studying the development and stability of complex, multifaceted constructs like social competence.

Focal Sampling

Focal person sampling involves selecting (typically at random from a roster of participants) one participant and observing the individual for a defined time period. For each sampling interval (ranges vary depending on the question of interest), the observer records all relevant behaviors of the focal person. As we have previously discussed (*see* Pellegrini et al., s in press), for studies of dyads or small groups, the sampling interval should be as long as the typical interaction or displayed behavior of interest. For example, in our work, we study the display of relational aggression (i.e., the use of the relationship as the means of harm via social exclusion, withdrawing friendship, spreading malicious rumors), and given the nature of these behaviors, we have found that an interval of 10 minutes is a reasonable interval for assessing the intent for harm as well as the subtle nature of these peer interactions (Ostrov, 2008; Ostrov & Keating, 2004).

Focal sampling may technically use continuous (e.g., Fagot & Hagan, 1985; Laursen & Hartup, 1989), 0/1 (e.g., Hall & McGregor, 2000; Harrist & Bradley, 2003), or instantaneous recording rules (*see* Pellegrini, 2004). However, focal sampling often uses continuous recording procedures because it permits the simultaneous coding of various behaviors, sequences of behaviors, and interactions with multiple partners in a live setting (e.g., Arsenio & Lover, 1997; Keating & Heltman, 1994). For example,

in our observational studies of relational aggression among young children, we always have used focal sampling with continuous recording given the somewhat covert nature of the behaviors we have targeted for observation, which require a longer period of direct assessment to decipher and appropriately record the behaviors (Ostrov & Keating, 2004). Focal participant sampling is often conducted across multiple days and contexts to better capture the true nature of the behavior rather than any state-dependent artifacts. Given the amount of time and the continuous nature of the recordings, this technique permits the recording of behavior that is a close approximation to real-time recording, and a researcher may recreate the behavior of the focal participants with a high degree of accuracy (Pellegrini et al., in press). For example, we observe children in their naturally occurring play contexts on 8 separate days, and they are only ever observed once per day to maintain independence of the data. Thus, in our work, each participant is observed for 80 minutes (8 sessions at 10 minutes each session). More specifically, a study of 120 children resulted in more than 370 hours of observation across the two time-points of the short-term longitudinal study (Ostrov, 2008). Therefore, time is a major cost of focal sampling because of the large number of independent observations typically conducted with this approach. Focal sampling may also be used with 0/1 or instantaneous sampling as recording procedures, but this is rarely done. As previously mentioned, both of these recording procedures require an *a priori* specified time interval, which is usually relatively brief (i.e., 1–10 seconds). Instantaneous recording is typically used only with scan sampling procedures (*see Scan Sampling* section below). 0/1 time sampling is not usually used with focal sampling because we are often interested in assessing the true frequency of behaviors that may not be obtained with this procedure (i.e., an independent behavior could occur once or more than once during a set interval, but with 0/1 coding only one point is scored).

Despite the emphasis on the use of these methods for studying basic social behavior, focal sampling procedures may be used in a wide range of studies. It is common in the literature to find focal participant sampling studies on a range of social behavior topics: social dominance in children (Keating & Heltman, 1994) and adults (Ostrov & Collins, 2007), play behavior (Pellegrini, 1989), emotion and aggression (Arsenio & Lover, 1997), conflict (Laursen & Hartup, 1989), and peer relations with young children and non-human primates (e.g., Hinde, Easton, & Meller, 1984; Silk, Cheney, & Seyfarth, 1996). However, there are many practical applications of focal participant sampling (*see* Leff & Lakin, 2005; Pellegrini, 2001). For example, applied studies have been conducted that have used these observational techniques for examining the adjustment of children with special needs in elementary schools (Hall & McGregor, 2000), peer victimization in early adolescence (Pellegrini & Bartini, 2000), and for testing the efficacy of randomized behavioral interventions (e.g., Harrist & Bradley, 2003; Ostrov et al., 2009).

Scan Sampling

Instantaneous or scan sampling is a more efficient observational procedure than focal sampling. Scan sampling exclusively relies on instantaneous recording rules (Pellegrini, 2001). With this procedure the observer scans the entire observation field for a possible behavior or event for a particular period of time. If an event is noted during that scan, then it is recorded. Typically, a number of discrete scans occur across a number of days to maximize the independence of the data. A participant's data is usually summed across the scans to yield a behavioral score for the construct of interest. A concern with this approach is that it may not accurately assess the true frequency of behaviors if spacing is not adequate between the scans (Pellegrini, 2004). Moreover, given the typical approach in which scans are conducted on an entire reference group in their natural context, behaviors that are selected for this approach must be readily apparent, discrete, and overt behaviors that require typically only a few seconds to observe. In our own field, McNeilly-Choque, Hart, Robinson, Nelson, and Olsen (1996) conducted a study of young children's aggressive behavior in which they used a random scan sampling method that yielded 100 five-second scans during a 5- to 7-week period, resulting in 8 minutes of total observation per participant (McNeilly-Choque, Hart, Robinson, Nelson, & Olsen 1996). Thus, this study demonstrated the feasibility and efficiency of systematic scan sampling observations of aggressive behavior on the playground.

Semi-Structured Observations

Analog tasks or semi-structured observations, involving controlled simulations or analog situations, are observational tasks designed to mimic naturalistic conditions. Semi-structured observational procedures are another observational paradigm well

suited for low base rate events. The recording and coding procedures are often identical to the procedures an observer would use in a naturalistic setting; however, the context in which the behaviors emerge is different. Often analog tasks are completed in a laboratory or similarly controlled setting and are videotaped for subsequent coding by unaware observers. Thus, analog observational paradigms permit a great deal of experimental control/standardization of procedures, and with the use of videotapes, observers are able to objectively code the session using the same recording rules as permitted in other contexts. A clear advantage of these procedures is that they are efficient and require less cost and time spent observing participants. If the study is not designed well, then a major disadvantage is a lack of ecological validity (i.e., degree to which the context in which the research is conducted parallels the real-life experience of the participants), and poor generalizability of the findings is possible. Moreover, a relatively small sampling of behavior does not provide for a true frequency of behavior or for a representative sample of behavior with many interaction partners (i.e., the researcher is not able to examine individual–partner interactions). Other researchers have addressed this concern by using a "round robin" approach in which each participant completes an analog session with several (or all) other member of the reference group, which may improve the validity of the approach but, of course, adds a great deal of time and expense (*see* Hawley & Little, 1999).

In our own research we have used a semi-structured observational paradigm to provide an efficient estimate of young children's aggressive behavior. To this end, we created a brief (9-minute) analog situation to observe various aggressive and prosocial behaviors (i.e., within dyads or triads) in early childhood (Ostrov & Keating, 2004; Ostrov, Woods, Jansen, Casas, & Crick, 2004). The procedures and a review of the psychometric findings are described extensively elsewhere (e.g., Ostrov & Godleski, 2007), but essentially, each assessment includes three trials of 3 minutes each. For each trial, the children are given the same developmentally appropriate picture to color (e.g., Winnie the Pooh). For triads, three crayons are placed on the table equidistant from all participants, and only one crayon is the functional instrument (e.g., orange crayon for Winnie the Pooh) and two are functionally useless white crayons. At the end of the trial, a new picture and new crayons are placed on the table. This procedure is designed to produce mild conflict among the children and was developed to permit the children to engage in a variety of behaviors: prosocial behavior (e.g., sharing the one functional crayon or breaking into pieces to share), relational aggression (e.g., telling the child they will not be their friend anymore unless they give them the crayon), and physical aggression (e.g., taking the crayon away from someone else). The analog task was designed to be developmentally appropriate and resemble everyday conflict interactions concerning limited resources that young children experience in their typical preschool classroom. Highly trained research assistants monitored the entire session and intervened if needed to guarantee the safety of all participants and reduce the likelihood of participant distress. Moreover, at the end of the session, the children were each individually given access to a full box of crayons to diminish any distress and they were praised for their performance (*see* Ostrov et al., 2004). This paradigm is thus designed to elicit the behavioral constructs of interest in a more controlled environment than free play yet ensures the ethical treatment of participants.

One way to demonstrate the ecological validity of semi-structured observations is to correlate behaviors observed in a semi-structured context with behaviors observed in a more naturalistic context. For example, Coie and Kupersmidt (1983) found that social status in experimentally contrived playgroups comprised of unfamiliar peers matched social status in the classroom, supporting the validity of a contrived playgroup paradigm for studying social development (*see also* Dodge, 1983). Similarly, our own brief semi-structured observational paradigm (i.e., coloring task) has been shown to significantly predict observational scores collected from concurrently assessed naturalistic (i.e., classroom and playground free play) focal child observations with continuous recording ($r = 0.48$) and to predict future (i.e., 12 months later) behavior in naturalistic contexts at moderate levels (*see* Ostrov et al., 2004).

Methods of Recording

Various methods of recording (i.e., checklist, detailed records, or observation forms) vary widely and should be based on the type of recording procedures that a researcher adopts. For example, time sampling (i.e., 0/1) and instantaneous or scan sampling procedures are well suited for checklist forms in which the prescribed intervals simply receive a check or a precise code indicating the occurrence

or absence of the behavior in question. However, focal participant sampling often requires observation forms that permit greater detail and several codes that are recorded either simultaneously or in close temporal proximity, and, as such, a form that includes the behaviors or events of interest with space for recording the behavior in detail may be needed (for example forms and templates, *see* Pellegrini et al., in press). A general concern here is that the more time spent writing details about the behavior/event removes the observer's attention from the participants and important details may be lost. Some observational procedures like time sampling provide the observer with a set period of time after the interval for recording behavior. In general, the easier the observation form is to complete, the less room there is for error. With that said, checklists often do not permit systematic reviews for accuracy of codes by the master trainer. For example, observers that are observing the same participant as part of a reliability check could both code a behavior as "PA" for physical aggression when in fact one research assistant observed a "hit" and the other observed a "kick," which, depending on the observational system, may be different and might not warrant a positive match or agreement. Thus, depending on the coding scheme and intentions of the researcher, these may artificially match for reliability purposes when in fact they were closely related but discrete behaviors. Finally, if observers record some written details about the event, they may inform subsequent decision rules concerning whether a recorded behavior from observer 1 matches or does not match observer 2 for reliability assessments.

Coding Considerations

The development of a reliable coding scheme is crucial for appropriately capturing the behaviors in question and testing the experimenter's *a priori* hypotheses (Bakeman & Gottman, 1987). There are three types of coding categories that are often included in observational systems: physical description codes, consequence codes, and relational or environmental relations codes (Pellegrini, 2004). Physical description is believed to be the most "objective" type of codes because these describe "muscle contraction" (Pellegrini, 2004, p. 108) and might, for example, be involved in recording a participant's social dominance or submissiveness (e.g., direct eye contact, rigid posture, arms akimbo; *see* Ostrov & Collins, 2007). The second type of codes is for those of consequence in which a constellation of behaviors are part of a single code if they lead to the same outcome (Pellegrini, 2004). For example, if we were interested in studying social dominance, then we might code taking objects away from others that result in a submissive posture on the part of the nonfocal participant to be an indicator of social dominance (Ostrov & Collins, 2007). The third type of codes includes categories in which participants are described in relation to the context in which they are observed (Pellegrini, 2004). An example of a relational observational category would be a coding scheme that accounted for where and with whom an individual was socially dominant. In terms of costs and benefits, it is clear that physical description codes are often easier to train and therefore potentially more reliable. It is possible that consequence codes may be unreliable given a misunderstanding of the sequence of events (Pellegrini, 2004). Relational codes involve the appropriate documentation of multiple factors and therefore create more possibilities of error (for discussion, *see* Pellegrini, 2004; Bakeman & Gottman, 1987). Overall, the level of analysis from micro- to macro-coding schemes is important to consider and the most objective and reliable system for addressing a researcher's particular research question should be adopted.

A second consideration is the determination of whether to use mutually exclusive and exhaustive codes. Mutually exclusive codes are used when a single behavior may be recorded under one and only one code. In our observational studies, our coding scheme includes mutually exclusive codes such that a single behavior may be coded as either physical aggression or relational aggression, but not both. Exhaustive coding schemes are designed such that for any given behavior of a theoretical construct, there is an appropriate code for that behavior. For example, in our work we have codes for physical, relational, verbal, or nonverbal aggression as well as aggression not otherwise specified. Thus, if we determine a behavior is an act of aggression, then it may be coded as one of our behaviors in our scheme. Often schemes include mutually exclusive and exhaustive codes because there are several benefits to this approach (*see* Bakeman & Gottman, 1987). Having mutually exclusive codes means that researchers are not violating assumptions of independence, which are often needed for parametric statistics. For example, if a single behavior may be coded as both physical and relational aggression, then that may violate our assumption that the data are independent and come from independent

behavioral interactions (Pellegrini, 2004). Having exhaustive codes also speaks to the content validity of a coding scheme. That is, if the overall construct appropriately measures all facets of that construct, then the behavior in question should be included in the observational system, and exhaustive schemes guarantee this occurrence. It is important to recall that the larger the coding scheme, the more taxing the observational procedures will be for observers and the greater the possibility of observer error.

Scoring

Scoring of observational data is similar to the scoring of any quantitative data within the social and behavioral sciences, and it often depends on the convention within a particular field and the type of observational sampling and recording techniques that are adopted. For example, for focal participant sampling with continuous recording, frequency counts are often generated by summing each independently recorded behavior across the various sessions. In our own research, that would mean that an individual participant would get a score for each of the constructs (i.e., physical aggression, relational aggression, verbal aggression, etc.) by summing all the behaviors within a construct (e.g., all physical aggression behaviors) across all eight sessions (Ostrov & Keating, 2004). If the number of sessions is different for each participant because of missing data, then it is often common practice to divide by the number of sessions completed to generate an average rate of behavior per session (*see* Crick, Ostrov, Burr et al., 2006). Occasionally it is apparent that an error was made in the original coding of behaviors. Best practices have not been established for addressing these concerns, but as long as these errors are not systematic, the adopted solutions are often not a concern. To avoid problems with potential scoring biases, the observers and coders should always be unaware of the participant's condition and/or past history. In addition, whenever possible, observers and coders should be unaware of the study hypotheses.

Psychometric Properties

Reliability

Reliability is often conceptualized as consistency within or between individuals (i.e., intra-observer or inter-observer), within measures (internal consistency), or across time (i.e., test–retest). Arguably, for observational methods, the most important measure of consistency is inter-observer reliability, or the degree to which two sets of observations from two independent observers agree (Stangor, 2011). In the present review, we will first address intra-observer reliability and then focus on the assessment of inter-observer reliability.

Intra-observer, or within-observer, reliability is defined as a situation in which two sets of observations by the same research assistant agree or are consistent. Essentially, intra-observer reliability is assessing how consistent a particular observer is when coding specific behaviors either between sessions (i.e., across time) or within a single session. As Pellegrini (2004) has discussed in more detail, we may conceptualize and test (e.g., Pearson's Product-Moment Correlation Coefficient) intra-observer reliability in ways similar to test–retest reliability, and thus, intra-observer reliability is essentially the temporal stability of the observational measure for a given observer between testing sessions. We might desire to know the degree to which the observational score on a given behavioral construct for the same observer is stable across time to test for observer drift (a threat to the validity of the observational data), or the likelihood that observers are deviating from initial training procedures over time and modifying the definitions of the constructs under study (Smith, 1986). Intra-observer reliability or consistency within an observer may also be conceptualized as the reliability of an observer's scores within a single session, and in this case the test is analogous to assessments of internal consistency (e.g., Cronbach's α). As Pellegrini (2004) has stated, we assume an observer is first reliable or consistent in their scoring/recording by themselves prior to testing if they agree with an independent observer (i.e., inter-observer reliability).

As mentioned, inter-observer reliability or consistency between observers is the gold standard for observational research. Essentially, inter-observer reliability involves comparing the independent codes of the observers with other trained observers. There are several ways to assess this psychometric property (*see* Pellegrini, 2004), but the key task is comparing agreement across all of the observers. An important best practice for inter-observer reliability procedures is to ensure that observers are sampling/recording the same behaviors independently. Independent coding may be conducted with the use of video and private coding sessions without discussion until all codes have been completed. Inter-observer reliability may be assessed live in the

field if the observers take precautions to avoid conveying to their partner how (and, in some cases, when) they are recording the behavior in question. A second best practice is to assess for reliability across the study to help avoid various biases (e.g., observer drift) and coding/recording errors from corrupting the integrity of the data. That is, observers should be checked against a master coder at the start of the study just after training ends, and each observer should pass an *a priori* reliability threshold (e.g., Cohen's $\kappa > 0.70$). Next, their observations should be compared against other independent reliable observers throughout the duration of the study, and the trainer should provide constructive feedback for any deviations from the training protocol. Finally, an important consideration is for what percentage of time inter-observer reliability will be checked. This percentage should be a function of the number of cases or possible events that will be recorded, but typically 15% to 30% of a randomly selected sample of the possible sessions is coded by more than one observer for assessing inter-observer reliability. To avoid potential biases, a best practice is for each observer to conduct reliability observations with all other observers in a round-robin format.

There are several ways to statistically measure inter-observer reliability. In the past, authors relied on zero-order correlations (Pearson's r) but that problematic practice is not seen as often in the recent literature. A second statistical method that is still reported in peer-reviewed journals is percent agreement. Percent agreement may be expressed in Equation 1:

$$P_{obs} = N_A/N_A + N_D \times 100\% \qquad (1)$$

where P_{obs} is the proportion of agreement observed, N_A is the total number of agreements, and N_D is the total number of disagreements. Percent agreement is not currently best practice, as it is influenced by the number of cases (i.e., it may be biased by relatively few cases) and because it is not compared against a standard threshold (Bakeman & Gottman, 1987). Finally, one of the central concerns with percent agreement (as well as Pearson's r) as a measure of inter-observer reliability is that it does not control for chance agreement (Bakeman & Gottman, 1987).

Cohen's (1960) κ is a preferred statistic for inter-observer reliability because it does control for chance agreements and is a more "stringent statistic," allowing greater precision in assessing reliability at a specific moment in time or for particular events rather than overall summaries of association (Bakeman & Gottman, 1987, p. 836). Importantly, κ may only be used when coders use a categorical scale (Bakeman & Gottman, 1987) and when a 2 x 2 matrix may be created to depict the proportion of agreements/disagreements for occurrences/nonoccurrences of behavior for any two observers (Pellegrini, 2004). When calculating the rate of agreement, it is important to *a priori* indicate any time parameters (i.e., within what period of time must both observers note the occurrence of a behavior, also known as the tolerance interval). Some experts caution that extremely short tolerance intervals (e.g., 1 sec) may be overly stringent and artificially reduce the degree of agreement given typical reaction times of observers (*see* Bakeman & Gnisci, 2006). If time sampling is being used, then observers should be signaled by an external source (e.g., audible tone from an electronic device) to indicate when they should record the behavior (*see* Pellegrini, 2004). κ may be expressed in Equation 2:

$$\kappa = P_{obs} - P_{exp}/1 - P_{exp} \qquad (2)$$

where P_{obs} is the proportion of agreement observed, and P_{exp} is the expected proportion of agreement by chance (Bakeman & Gnisci, 2006). Equation 2 indicates that agreement anticipated as a result of chance is subtracted from both the numerator and denominator, thus κ provides the proportion of agreement corrected for chance agreements (Bakeman & Gnisci, 2006). The range for κ is from -1.00 to $+1.00$, with a value of "0" indicating that obtained agreement is equivalent to agreement anticipated by chance, and greater than chance agreement would yield positive values with +1.00 equal to perfect agreement between the observers (Cohen, 1960). Interestingly, Cohen (1960) revealed that negative values (less than 0) were rare and suggested agreement at less than chance levels. It is possible to test if κ is significantly different from 0, but statistical significance is often not used as a threshold for determining an "adequate" or "good" criterion (Bakeman & Gottman, 1987). Initially, Landis and Koch (1977) provided an index of the strength of agreement or "benchmarks" and reported the following standards: κ of < 0.00 was "poor," $0.00 - 0.20$ was "slight," $0.21 - 0.40$ was "fair," $0.41 - 0.60$ was "moderate," $0.61 - 0.80$ was "substantial," and > 0.81 was "almost perfect" (p. 165). However, Bakeman and Gottman (1987) reported that a significant κ of less than 0.70 may be a reason for concern. Other scholars have noted that the conservative nature of κ permits one to use a slightly lower threshold for adequate levels of reliability than the

typical convention of 0.70 and suggest that a κ coefficient of 0.60 or higher is "acceptable" and 0.80 or above is considered "good" (Pellegrini, 2001).

Under circumstances when a κ coefficient may not be calculated (e.g., when noncategorical data is used or quadrants of the aforementioned occurrence matrix may not be available given the recording rules of the adopted observational procedure), scholars have suggested that an intraclass correlation coefficient (ICC) be computed between independent raters on the continuous data (Bartko, 1976; McGraw & Wong, 1996; Shrout & Fleiss, 1979). There are several possible ICC formulas that could be depicted that are beyond the scope of the present review, and as such the interested reader is referred to the prior literature on this topic (Shrout & Fleiss, 1979; McGraw & Wong, 1996). Intra-class correlation coefficients may be expressed as a function of either the reliability for a single rating (i.e., the reliability of a typical, single observer compared to another observer) or the average rating of the observations across all the raters (McGraw & Wong, 1996). The average rating ICC uses the Spearman-Brown correction to indicate the reliability for all the observers averaged together (Bartko, 1976). The absolute value of an ICC assessing average ratings will be greater or equal to the ICC for a single rater (Bartko, 1976). Intra-class correlation coefficients may also be calculated as an index of "consistency" or as a measure of "absolute agreement." Essentially, if systematic differences among observers are of interest, then the "absolute agreement" formula accounts for observer variability in the denominator of the ICC estimate, and this is not included for ICCs that measure "consistency" (for further detail, *see* McGraw & Wong, 1996). Intra-class correlation coefficients range from –1.00 to +1.00, where negative values indicate a lack of reliability and +1.00 would indicate perfect agreement (Bartko, 1976). An advantage to ICCs is that confidence intervals may be calculated (*see* McGraw & Wong, 1996). Typically, acceptable levels of reliability for ICCs are similar to other criteria in the field, and as such, levels greater than or equal to 0.70 are considered "acceptable" (e.g., Ostrov, 2008; NICHD Early Child Care Research Network, 2004).

Validity

In using observational research methods, an assessment of validity is equally as important as an assessment of reliability. Different types of validity should be considered to strengthen the inferences drawn from a particular method, with construct validity being most fundamental to any empirical inquiry. Construct validity is the degree to which the construct being studied actually measures the concept that a researcher intends to study (Stangor, 2011). Construct validity is often established through assessments designed to measure convergent and discriminant validity. Convergent validity rests on the assumption that if a construct is truly being measured, then alternative assessments of the same construct should be correlated with each other (Stangor, 2011). For example, an observational method intended to measure disruptive behaviors in the classroom should be correlated with teacher reports of disruptive behaviors. Alternatively, discriminant validity suggests that the construct being studied should not be correlated with other variables unrelated to the construct (Stangor, 2011). Should the expected convergent and discriminant associations not be observed, then it is unclear what an instrument or observational system is measuring.

Other types of validity that are secondary yet still important to the establishment of a psychometrically sound observational system include content validity and criterion validity. Content validity refers to the extent to which a measure adequately assesses the full breadth of the construct being studied (Stangor, 2011). For example, an observational study of children's play behavior should code for different types of play, given that it is a diverse construct. To ensure that all facets of a construct are included in an observational system, correspondence with experts and focus groups/review panels may be used. Criterion validity involves an assessment of whether a study variable is associated with a theoretically relevant outcome measure. If observations are associated with an outcome that is measured at the same point in time at which observations are conducted, then concurrent validity is demonstrated. If observations are associated with an outcome that is measured at a future point in time, then predictive validity is demonstrated. For example, concurrent validity would be confirmed by associations between classroom observations of disruptive behavior and teacher report of rejection by peers, and predictive validity would be confirmed by associations between classroom observations of disruptive behavior and future parent -report of academic performance.

THREATS TO VALIDITY: SOURCES OF BIAS AND ERROR

There are numerous biases for which observational methods are susceptible. A key bias is the

aforementioned observer drift, and it is paramount that investigators monitor for this threat to the validity of the data by carefully assessing observational records and calculating reliability coefficients for the duration of the study. Importantly, in addition to the aforementioned discussion about intra-observer reliability, observer drift may also be indicated if there is a drop in inter-observer reliability among the phases of training and data collection (Smith, 1986). A second strategy to mitigate observer drift is to regularly retrain observers. In instances where particular observers demonstrate problematic coding patterns, retraining should be individualized and should target the particular area of concern. In general, retraining is a practice that is beneficial for every observer because it reinforces proper coding procedures and observer behavior, thereby ensuring the integrity of the study.

A second type of distortion that must be considered results from participant reactivity, which is also a threat to the validity of the observational data. Reactivity occurs when the individuals under study alter their behavior because of the presence or influence of an observer. Consequently, the behavior observed does not provide a true representation of the construct being measured. If participants avoid a particular location within a setting or modify their behavior because they know they are being recorded, this is a major concern for the validity of the data (Stangor, 2011). Depending on the nature of the study, reactivity may be more probable. For example, when observers need to remain within earshot of a focal participant to hear and see the behavioral interactions, it is crucial that the observers remain unobtrusive (e.g., Pellegrini, 1989). Researchers should explicitly address reactivity by training observers in the field to have a minimally responsive manner (Pellegrini, 2004). Essentially, observers should use neutral facial expressions and control their nonverbal behavior, posture, movement, and reactions to events during live coding. It is also possible that participants may be reactive to cameras and other recording devices, and efforts should be made to habituate participants to this equipment (*see Use of Technology and Software* section below) and monitor for this occurrence. Thus, this habituation process should occur prior to the actual collection of data (Pellegrini, 2004). In our studies, we spend a minimum of several days in the observational environment (and will do so for as long as needed) simulating our observations, which provide the participants an opportunity to habituate to our presence and reduce reactivity prior to actual data collection. Therefore, regardless of live or videotaped coding, researchers should observe for participant reactivity and report the degree of reactivity in their studies (e.g., Atlas & Pepler, 1998). We define participant reactivity as any direct eye contact between the focal participant and observer, comments from the focal participant to the observer about our presence, or comments about our presence to others in the environment (Ostrov, 2008). Our training procedures and careful monitoring has resulted in relatively low levels of reactivity in several studies (e.g., 1.5–2.5 times per focal participant during 80 min of observation; Crick, Ostrov, Burr et al., 2006).

Observer expectancy effects are a third bias (Hartmann & Pelzel, 2005), which is essentially when observers form expectations about the nature of the data based on their knowledge or assumptions about the study goals and hypotheses, which is why best practice is to use unaware observers, when possible, and to use unaware observers for reliability purposes, at a minimum.

A final source of bias that we will discuss is gender bias as this is a well-documented concern with observational methods (Ostrov, Crick, & Keating, 2005). Past research has documented that untrained observers maintain gender biases when observing, for example, physical aggression (Lyons & Serbin, 1986; *see also* Condry & Ross; 1985; Susser & Keating, 1990). That is, men tend to rate boys as more physically aggressive than girls, even when boys and girls are displaying comparable levels of aggression (Lyons & Serbin, 1986). Moreover, male and female college students have shown documented gender biases based on knowledge about gender of young children in past experimental studies (Gurwtiz & Dodge, 1975). Finally, in our own research, we have documented that male college students are less likely to correctly identify relational aggression or prosocial behavior than their female peers (Ostrov et al., 2005). Please note that although the examples were related to our field of study (i.e., aggression), gender biases may be present for a variety of topics of study. Importantly, it may be that when individuals are trained to recognize potential biases, they are more likely to be objective in their coding of behavior (Lyons & Serbin, 1986).

Use of Technology and Software

Excellent detailed reviews of computer-assisted recording devices and observational software programs are available (*see* Hoch & Symons, 2004),

and thus, the present goal of this section is to briefly review the current state of technology and software for assisting in systematic observations in the laboratory and field. The following will include a review of the three most common observational software programs as well as the use of handheld devices and remote audiovisual equipment. The commercially available programs vary widely in function and cost, but most permit the observer to define a coding scheme and corresponding letter or number codes that observers can quickly use when making observations live or when coding digital media in the laboratory. Overall, advances in technology have made observational methods more efficient (e.g., flexible data reduction procedures and automatic statistical analyses), accurate (i.e., automatic rewind and playback functions reduce errors in coding), and applicable to a wider range of settings and topics of study (Bakeman & Gnisci, 2006, p. 140).

The first software program and associated computer-assisted recording devices that we will discuss is the Observer® system by Noldus Inc. (Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000). The current version is Observer XT, which permits both time sampling as well as continuous event-based observational systems and has been used in both human and animal research (*see* http://www.noldus.com/the-observer-xt/observer-xt-research). A notable feature is that this software permits an assessment of response latency of the time between the onset of a stimulus and the initiation of the response, which facilitates consequence coding (*see Coding Considerations* section above). The software also permits the linking of data from multiple modalities (e.g., observational reports, physiological responses) with a continuous time synch. The software may be used in the field with durable handheld devices or in the laboratory with live streaming video linked directly with the coding program (Noldus et al., 2000). Finally, the new version of the software permits searches of the data for particular comments, events, or behaviors, and data may be exported to various statistical software packages (Noldus et al., 2000). Jonge, Kemner, Naber, and van Engeland (2009) used an earlier version of the Observer software to code data from a study on block design reconstruction in children with autism spectrum disorders and a group of comparison participants. The use of the videotaped sessions and later coding by unaware observers meant that the coders using the software were unaware of the child's group status. The software permitted the coders to record the amount of time the children took to reconstruct the block design pattern as well as a range of errors (Jonge et al., 2009). The program was used to calculate Cohen's κ based on two independent coders (Jonge et al., 2009), who could make independent evaluations of the behavior without biasing their coding partner.

The second observational software program that we examine is the Multi-Option Observation System for Experimental Studies (MOOSES; Tapp, Wehby, & Ellis, 1995) and the associated Procoder for Digital Video (PCDV; Tapp& Walden, 1993), which permits viewing and coding of digital media (*see* http://mooses.vueinnovations.com/overview). The MOOSES and PCDV programs also permit event and time sampling and for the coding of real-time digital media files or verbatim transcripts of observational sessions (Tapp & Walden, 1993; Tapp et al., 1995). In fact, data files may be exported to MOOSES for event coding or to another format known as the Systematic Analysis of Language Transcripts (SALT) for transcription data coding. MOOSES automatically timestamps events and may provide frequency and duration codes as well as basic reliability statistics (e.g., Cohen's κ), and MOOSES is designed for sequential analysis (Tapp et al., 1995). A handheld version of MOOSES is available. MOOSES/PCDV has been described as a lower cost alternative to The Observer (Hoch & Symons, 2004).

The third system we review is the Behavior Evaluation Strategies and Taxonomies (BEST; Sharpe & Koperwas, 2003). This computer system includes both the *BEST Collection* for capturing digital media files and the *BEST Analysis* program for both qualitative and quantitative analysis of the observational data (Sidener, Shabani, & Carr, 2004). The BEST program may be used for examining the frequency or duration of events, and sophisticated sequential analysis may be conducted. Much like the more expensive alternatives, this program will calculate reliability statistics (e.g., Cohen's κ) and will summarize data in table or various graph formats. A review of this program suggests that BEST does not handle the collection of interval-based data well, but the BEST Analysis program will allow a researcher to analyze this type of observational data (Sidener et al., 2004). A new platform permits video display for captured data from video files, and although the program was initially written for Windows®, there are inexpensive Apple® iPhone® and iPod Touch® applications available for data collection (*see* http://www.skware.com).

Various types of technology (e.g., audio and video recordings) have an extensive history in the field and laboratory to assist researchers in better capturing verbal and nonverbal interactions (e.g., Abramovitch, Corter, Pepler, & Stanhope, 1986; Stauffacher & DeHart, 2005). Remote audiovisual recordings provided an opportunity to combine the benefits of both audio and video recording while also reducing reactivity to typical recording devices when participants were observed in naturally occurring settings (Asher & Gabriel, 1993; Atlas & Pepler 1998; Pellegrini, 2004; Pepler & Craig, 1995; Pepler, Craig, & Roberts, 1998). That is, videotaping with a telephoto zoom lens from an unobtrusive location in the natural setting and recording audio via a system of wireless microphones provides an externally valid way to record behavior and a time-synched verbal record of the interaction (Pepler & Craig, 1995). Thus, remote audiovisual observational recordings provide all the benefits of having a video for subsequent coding by unaware observers (i.e., the ability to pause, rewind, and analyze subtle nonverbal behaviors) as well as a complete verbal transcript, which helps to put the video data in proper context (Asher & Gabriel, 1993; Pepler & Craig, 1995). Wireless microphones typically are housed within small vests or waist pouches that participants wear, and often only the focal participant has an active or live microphone, and others in the reference group have "dummy" microphones that resemble the weight and look of the real microphone. Importantly, observational codes made with the remote audiovisual equipment have demonstrated acceptable inter-observer reliability coefficients (e.g., $\kappa = 0.76$; Pepler & Craig, 1995). Moreover, this procedure as well as sufficient exposure to the equipment by the participants has been found to produce low levels of participant reactivity (e.g., <5% , Atlas & Pepler, 1998; *see also* Asher & Gabriel, 1993). The benefits of a rich observational record with low levels of reactivity within settings of high ecological validity seem to outweigh the costs, which include additional training, equipment costs, and some ethical considerations. A central ethical consideration is that individuals without consent may be recorded indirectly. A possible solution is to temporarily store and then, after processing, discard film clips of individuals without consent (Pepler & Craig, 1995), but this solution may violate the rights of nonparticipants. Alternatively, a researcher could restrict access to the observational setting to only those with consent, but this second approach is a threat to the ecological validity of the procedures (Pepler & Craig, 1995). An additional concern is that third parties may wish to use the data as surveillance, which might limit the rights of participants being recorded. As such, policies related to confidentiality and any possible limits of confidentiality should be discussed with the participants and any other possible party that may desire access to the data (*see* Pepler & Craig, 1995). Importantly, to our knowledge, remote audiovisual observational methodology has only been used with school-aged children in the classroom (Atlas & Pepler, 1998) and typically on the playground (e.g., Asher & Gabriel, 1993; Pepler, Craig, & Roberts, 1998); thus, it is not clear if older individuals would be more aware and reactive to the procedure and equipment (Pepler & Craig, 1995).

Ethical Considerations

There are several ethical considerations with observational research. With naturally occurring phenomena, there may be a temptation to observe social interactions and behavior without obtaining informed consent. Although this practice may technically be exempt from most Institutional Review Board (IRB) review (i.e., if identifying information is not collected and video or audio recordings of the public behavior are not made), we strongly encourage researchers to obtain informed consent from participants and assent from legal minors to support their right for autonomy but also so that all risks (e.g., breaches of confidentiality) may be appropriately conveyed. To avoid these breaches of confidentiality, researchers conducting live observations typically use identification codes rather than identifying information about the participants on all observation forms and in data files. Access to video or audio recordings of observational sessions is typically restricted to only those individuals (e.g., coders) who must have access as part of the research study. Participants should be fully informed for how long the observational recordings will be maintained and when they will be destroyed. A final ethical consideration concerns intervention efforts or at what point the researcher or observers will intervene (for a discussion of duty to warn with observational methods, *see* Pepler & Craig, 1995) and directly or indirectly act on the behalf of the participants. For example, in our observational studies, we have clearly established procedures for when we will notify a teacher that a child in the observation setting is in danger or in need of help (e.g., leaving the controlled area, serious injury). These

procedures are discussed at the start of the study with school officials and are part of our consent process, which we believe are best practices.

An Overview of Procedures for a High-Quality Systematic Observational Study

The researcher begins by *a priori* selecting and operationally defining behaviors of interest. Next, the researcher adopts a coding scheme by selecting the most appropriate sampling and recording procedures given the nature of the behavior under study and the observational context (*see* Table 15.2). Ethical considerations should be addressed during this development stage of the observational method and should be evaluated for the duration of the study. If the observational scheme is newly developed for the study, then it is imperative that pilot testing occur within a similar context and with a sample representing the target population. If it is not a new scheme or if pilot testing does not indicate any problems, then the investigator may begin training observers. If there are problems noted, then it is important to rectify these issues as quickly as possible to avoid further errors in the study. It is possible that modifications will be needed regarding the operational definition of the observed constructs or changes may be needed to the procedures and coding scheme given the nature of the context or sample under study. Once these changes are adopted, additional checks should be made to verify the solution has worked to ameliorate the original concerns. Training involves the use of a standardized manual, and initial reliability training assessments are conducted prior to the collection of data. Behavior is sampled in the lab or in the field in accordance with the adopted sampling and recording rules, and interobserver reliability is collected for the duration of the study. Validity assessments are also conducted using alternative informants and methods. If reliability or validity problems are detected, then this may also yield further modifications to the coding scheme to address the problems. If no psychometric problems are noted, then coding and scoring of the observational data occurs using standardized procedures. Finally, the data are analyzed and reported, which concludes the systematic observational study (*see* Fig. 15.1).

Conclusion

Systematic observational methods provide an opportunity to record the behavior of humans and animals in a relatively objective manner, without sacrificing ecological validity. In the present chapter, we have attempted to identify best practices as well as benefits and costs of various sampling and recording techniques. Quantitative researchers should be guided by *a priori* research questions and hypotheses

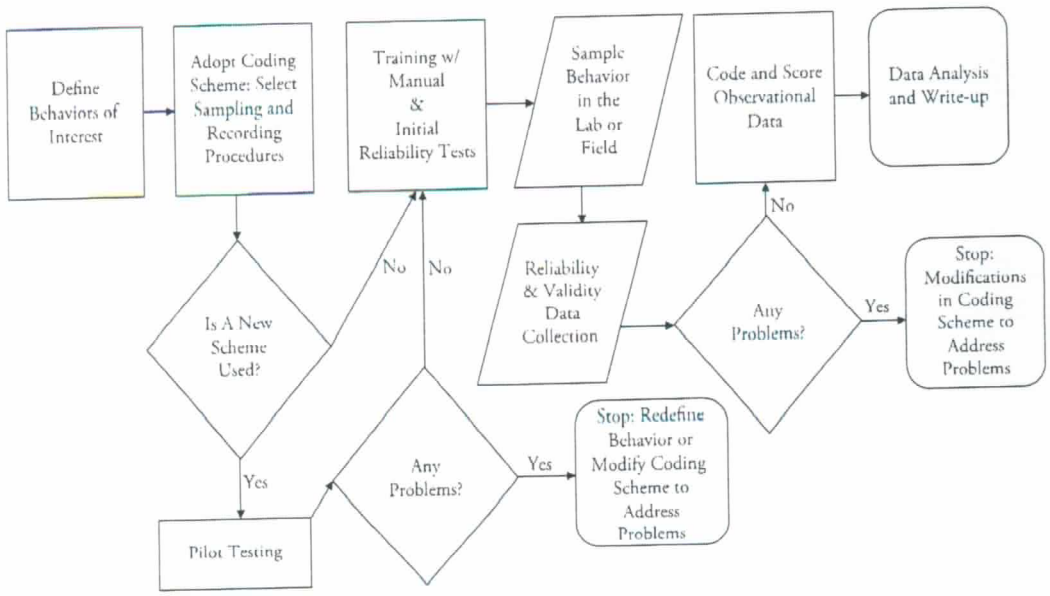

Figure 15.1 Procedures for a high-quality systematic observational study.

when selecting the most appropriate sampling and recording procedure for the specific research setting. Systematic observations require careful attention to coding and scoring decisions and a focus on achieving acceptable levels of reliability and validity. As a field, we must work to establish more stringent standards of reliability (i.e., inter-observer) and validity (i.e., construct) for observational methods. Moreover, we must continue to address and reduce various sources of bias and error. The use of computer-assisted software and digital analysis technology provide some promising options for increasing the efficiency and appeal of systematic observations in the field. Attention must also be given to key ethical considerations to guide appropriate conduct as an observational researcher. Careful consideration of these issues may inform quality research in a wide variety of basic, clinical, and educational contexts.

Future Directions

Observational methods have been a part of the social and behavioral sciences since the early years of our field, and we anticipate that there is a bright future for observational methods within the quantitative scholar's toolbox. We have defined seven questions and two remaining issues that we believe the field should work to address. This list is not exhaustive, but we hope these questions will generate future work using systematic observational methods.

1. What is the utility of observational methods above and beyond additional informants? Given the time and cost of observational methods, it is necessary to continue to demonstrate that observational methods have incremental predictive utility or may explain unique amounts of variance in relevant outcomes, above and beyond other informants and measures (Doctoroff & Arnold, 2004; Shaw et al., 1998). For example, we have demonstrated that observations of relational and physical aggression account for a significant amount of unique variance above and beyond teacher reports of relational and physical aggression in the prediction of teacher-reported deceptive and lying behaviors (Ostrov, Ries, Stauffacher, Godleski, & Mullins, 2008).

2. How does one best examine the construct validity of observational methods? To date, there is not wide consensus on the best approach for demonstrating the construct validity of observational systems. The typical approach is to compare observational data to other "gold standard" methods. For example, convergent evidence is achieved when high levels of association are found across methods such as between observations of aggression subtypes in classrooms, observations of aggression subtypes via semi-structured observations, and with various informants including teacher reports and parent reports of aggression subtypes (e.g., Crick, Ostrov, Burr, et al., 2006; Hinde et al., 1984; Ostrov & Bishop, 2008; Ostrov & Keating, 2004; Pellegrini & Bartini, 2000).

3. How do we detect observer biases? We believe the field has only begun to address the important issue of how to assess and identify observer biases. Much further work is needed to examine a host of possible biases from observer drift and observer expectancy effects to gender biases as well as other possible sources of distortion such as halo effects and potential expectancy biases derived from prior knowledge of participants in longitudinal studies (Hartmann & Pelzel, 2005). In addition, more focus should be placed on assessing participant reactivity. Few studies report this source of error and threat to validity, and we encourage observational researchers to quantify the degree to which their participants are reactive to the observational procedures.

4. How do we eliminate observer biases and other sources of error? Once we identify observer biases, we need more evidence-based information on how to appropriately eliminate these biases and sources of error. The literature has indicated few possible solutions (e.g., increased training for individuals with identified biases). In addition, more emphasis should be placed on identifying best practices for reducing reactivity. It is clear that minimally responsive procedures and habituation practices have worked effectively to reduce reactivity to low levels (e.g., <5% of time), but our goal should be to eliminate this source of error from our data.

5. What is the sufficient amount of time for observational sampling? Too often the time interval for time sampling as well as the total duration of observed time for event-based coding systems is decided without sufficient justification, and greater work is needed to establish parameters and strategies for determining the most efficient and useful time intervals for various behaviors and settings.

6. How do we reduce the cost of observational methods? One of the biggest obstacles to greater adoption of systematic observational methods is the cost of observational procedures. Typically, large staffs of highly trained individuals are needed for observational work, and although volunteer research assistants may be used to address this concern, this is still a significant barrier to further work in this area. Moreover, the overall amount of time to conduct an observational study is potentially longer than comparable studies with other methods, and thus we must work to make training procedures, data collection, and coding processes more efficient. The use of computer-assisted software and coding technology will continue to greatly help in this regard.

7. How do we refine and create observational software so that it is compatible with all types of observational systems and more flexible as well as affordable? Although observational software and recording devices have advanced a great deal in recent years (*see* Hoch & Symons, 2004), the software must become more flexible to accommodate a greater range of observational sampling and recording procedures. Moreover, the financial cost of these programs and licenses are often prohibitive, and efforts must be made to develop high-quality, affordable, and flexible computer-assisted observational software programs.

8. A key remaining issue is that as a field we need to move away from the use of Pearson product moment correlations and percent agreement as a standard measure of assessing inter-observer reliability. Given what we know about the role of chance agreement from classic (e.g., Cohen, 1960) and modern sources (Bakeman & Gottman, 1987; Pellegrini, 2004), it is not clear why some peer-reviewed manuscripts continue to only present either Pearson product moment correlations or percent agreement as strong evidence of inter-observer reliability.

9. A second remaining concern is that greater discussion of the ethical issues involved in observational methods is needed. For example, as we have discussed, it is not always clear when intervention is needed by observers in the field. Further, greater work needs to be conducted to examine how we may best ensure confidentiality of data with detailed observational records. Finally, we must focus on how we ensure confidentiality with the transfer of electronic observational data via handheld devices and other electronic technology.

Author Note

We wish to thank Jennifer Kane and members of the UB Social Development Laboratory for their assistance with the preparation of this chapter. Thanks to Dr. Leonard J. Simms for comments on an earlier draft. Special thanks to Dr. Anthony D. Pellegrini, who has greatly influenced the way we conceptualize systematic observational methods. Please direct correspondence to the first author at jostrov@buffalo.edu or 716-645-3680.

References

Abramovitch, R., Corter, C., Pepler, D. J., & Stanhope, L. (1986). Sibling and peer interaction: A final follow-up and a comparison. *Child Development, 57*, 217–229.

Arrington, R. E. (1943). Time sampling in studies of social behavior: A critical review of techniques and results with research suggestions. *Psychological Bulletin, 40*, 81–124.

Arsenio, W. F., & Lover, A. (1997). Emotions, conflicts and aggression during preschoolers' free play. *British Journal of Developmental Psychology, 15*, 531–542.

Asher, S. R., & Gabriel, S. W. (1993). Using a wireless transmission system to observe conversation and social interaction on the playground. In C. H. Hart (Ed.). *Children on playgrounds: Research perspectives and applications* (pp. 184–209). Albany, NY: SUNY Press.

Atlas, R. S., & Pepler, D. J. (1998). Observations of bullying in the classroom. *Journal of Educational Research, 92*, 86–99.

Bakeman, R., & Gnisci, A. (2006). Sequential observational methods. In M. Eid & E. Diener (Eds.). *Handbook of multimethod measurement in psychology* (pp. 127–140). Washington DC: American Psychological Association.

Bakeman, R., & Gottman, J. M. (1987). Applying observational methods: A systematic view. In J. D. Osofsky (Ed.). *Handbook of infant development.* (2nd ed., pp. 818–854). New York: John Wiley.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*, 762–765.

Bowker, A., Boekhoven, B., Nolan, A., Bauhaus, S., Glover, P, Powell, T., & Taylor, S. (2009). Naturalistic observations of spectator behavior at youth hockey games. *Applied Research, 23*, 301–316.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Coie, J. D. & Kupersmidt, J. B. (1983). A behavioral analysis of emerging social status in boys' groups. *Child Development, 54*, 1400–1416.

Condry, J. C., & Ross, D. F. (1985). Sex and aggression: The influence of gender label on the perception of aggression in children. *Child Development, 56*, 225–233.

Crawford, M. P. (1942). Dominance and social behavior, for chimpanzees, in a non-competitive situation. *Journal of Comparative Psychology, 33*, 267–277.

Crick, N. R., Ostrov, J. M., Burr, J. E., Jansen-Yeh, E. A., Cullerton-Sen, C., & Ralston, P. (2006). A longitudinal study of relational and physical aggression in preschool. *Journal of Applied Developmental Psychology, 27*, 254–268.

Doctoroff, G. L., & Arnold, D. H. (2004). Parent-rated externalizing behavior in preschoolers: The predictive utility of structured interviews, teacher reports, and classroom observations. *Journal of Clinical Child and Adolescent Psychology, 4*, 813–818.

Dodge, K. A. (1983). Behavioral antecedents of peer social status. *Child Development, 54*, 1386–1399.

Englund, M. M., Levy, A. K., Hyson, D. M., & Sroufe, L. A. (2000). Adolescent social competence: Effectiveness in a group setting. *Child Development, 71*, 1049–1060.

Fagot, B. T., & Hagan, R. (1985). Aggression in toddlers: Responses to the assertive acts of boys and girls. *Sex Roles, 12*, 341–351.

Goodenough, F. L. (1930). Inter-relationships in the behavior of young children. *Child Development, 1*, 29–47.

Gurwtiz, S. B., & Dodge, K. A. (1975). Adults' evaluations of a child as a function of sex of adult and sex of child. *Journal of Personality and Social Psychology, 32*, 822–828.

Hall, L. J., & McGregor, J. A. (2000). A follow-up study of the peer relationships of children with disabilities in an inclusive school. *The Journal of Special Education, 34*, 114–126.

Harrist, A. W., & Bradley, K. D. (2003). You can't say you can't play: Intervening in the process of social exclusion in the kindergarten classroom. *Early Childhood Research Quarterly, 18*, 185–205.

Hartmann, D. P., & Pelzel, K. E. (2005). Design, measurement, and analysis in developmental research. In M. H. Bornstein & M. E. Lamb (Eds.). *Developmental science: An advanced textbook* (5th ed., pp. 103–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Hawley, P. H., & Little, T. D. (1999). On winning some and losing some: A social relations approach to social dominance in toddlers. *Merrill-Palmer Quarterly, 45*, 185–214.

Hinde, R. A., Easton, D. F., & Meller, R. E. (1984). Teacher questionnaire compared with observational data on effects of sex and sibling status on preschool behavior. *Journal of Child Psychology and Psychiatry, 25*, 285–303.

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.). *Best practices in school psychology-IV* (pp. 993–1006). Bethesda, MD: National Association of School Psychologists.

Hoch, J., & Symons, F. J. (2004). Computer-assisted recording and observational software programs. In A. D. Pellegrini's *Observing children in their natural worlds: A methodological primer*. (2nd ed., pp. 214–222). Mahwah, NJ: Lawrence Erlbaum Associates

Jackson, H. G. & Neel, R. S. (2006). Observing mathematics: Do students with EBD have access to standards-based mathematics instruction? *Education and Treatment of Children, 29*, 593–614.

Jonge, M. de., Kemner, C., Naber, F., & Engeland, H. van. (2009). Block design reconstruction skills: not a good candidate for an endophenotypic marker in autism research. *European Child & Adolescent Psychiatry, 18*, 197–205.

Keating, C. F., & Heltman, K. R. (1994). Dominance and deception in children and adults: Are leaders the best misleaders? *Personality and Social Psychology Bulletin, 20*, 312–321.

Krehbiel, D., & Lewis, P. T. (1994). An observational emphasis in undergraduate psychology laboratories. *Teaching of Psychology, 21*, 45–48.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Langfeld, H. S. (1913). Text-books and general treatises. *Psychological Bulletin, 10*, 25–32.

Länsisalmi, H., Peiró, J. M., & Kivimäki, M. (2000). Collective stress and coping in the context of organizational culture. *European Journal of Work and Organizational Psychology, 9*, 527–559.

Laursen, B., & Hartup, W. W. (1989). The dynamics of preschool children's conflicts. *Merrill-Palmer Quarterly, 35*, 281–297.

Leff, S.S., & Lakin, R. (2005). Playground-based observational systems: A review and implications for practitioners and researchers. *School Psychology Review, 34*(4), 475–489.

Lopez-Williams, A., Chacko, A., Wymbs, B. T., Fabiano, G. A., Seymour, K. E., Gnagy, E. M., et al. (2005). Athletic performance and social behavior as predictors of peer acceptance in children diagnosed with attention-deficit/hyperactivity disorder. *Journal of Emotional and Behavioral Disorders, 13*, 172–180.

Lyons, J. A., & Serbin, L. A. (1986). Observer bias in scoring boys' and girls' aggression. *Sex Roles, 14*, 301–313.

Macintosh, K. & Dissanayake, C. (2006). A comparative study of the spontaneous social interactions of children with high-functioning autism and children with Asperger's disorder. *Autism, 10*, 199–220.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intra-class correlation coefficients. *Psychological Methods, 1*, 30–46.

McNeilly-Choque, M. K., Hart, C. H., & Robinson, C. C., Nelson, L., & Olsen, S. F. (1996). Overt and relational aggression on the playground: Correspondence among different informants. *Journal of Research in Childhood Education, 11*, 47–67.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of applied behavior analysis, 40*, 501–514.

Newcomb, T. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology, 22*, 279–289.

NICHD Early Child Care Research Network. (2004). Trajectories of physical aggression from toddlerhood to middle childhood. *Monographs of the Society for Research in Child Development, 69*, (*Serial No.* 278).

Noldus, L. P., Trienes, R. J., Hendriksen, A. H., Jansen, H., & Jansen, R. G. (2000). The observer video-pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, and Computers, 32*, 197–206.

Ostrov, J. M. (2008). Forms of aggression and peer victimization during early childhood: A short-term longitudinal study. *Journal of Abnormal Child Psychology, 36*, 311–322.

Ostrov, J. M. & Bishop, C. M. (2008). Preschoolers' aggression and parent-child conflict: A multiinformant and multimethod study. *Journal of Experimental Child Psychology, 99*, 309–322.

Ostrov, J. M., & Collins, W. A. (2007). Social dominance in romantic relationships: A Prospective longitudinal study of non-verbal processes. *Social Development, 16*, 580–595.

Ostrov, J. M., Crick, N. R., & Keating, C. F. (2005). Gender-biased perceptions of preschoolers' behavior: How much is aggression and prosocial behavior in the eye of the beholder? *Sex Roles, 52*, 393–398.

Ostrov, J. M., & Godleski, S. A. (2007). Relational aggression, victimization, and language development: Implications for practice. *Topics in Language Disorders, 27*, 146–166.

Ostrov, J. M., & Keating, C. F. (2004). Gender differences in preschool aggression during free play and structured interactions: An observational study. *Social Development, 13*, 255–277.

Ostrov, J. M., Massetti, G. M., Stauffacher, K., Godleski, S. A., Hart, K. C., Karch, K. M., Mullins, A. D., et al. (2009). An intervention for relational and physical aggression in early childhood: A preliminary study. *Early Childhood Research Quarterly, 24*, 15–28.

Ostrov, J. M., Ries, E. E., Stauffacher, K., Godleski, S. A., & Mullins, A. D. (2008). Relational aggression, physical aggression and deception during early childhood: A multi-method, multi-informant short-term longitudinal study. *Journal of Clinical Child and Adolescent Psychology, 37*, 664–675.

Ostrov, J. M., Woods, K. E., Jansen, E. A., Casas, J. F., & Crick, N. R. (2004). An observational study of delivered and received aggression, gender, and social-psychological adjustment in preschool: "This white crayon doesn't work. . ." *Early Childhood Research Quarterly, 19*, 355–371.

Parten, M. B. (1932). Social participation among pre-school children. *The Journal of Abnormal and Social Psychology, 27*, 243–269.

Pelham, W. E. Jr., Gnagy, E. M., Greiner, A. R., Hoza, B., Hinshaw, S.P. Swanson, J. M., et al. (2000). Behavioral versus behavioral and pharmacological treatment in ADHD children attending a summer treatment program. *Journal of Abnormal Child Psychology, 28*, 507–525.

Pellegrini, A. D. (1989). Categorizing children's rough-and-tumble play. *Play & Culture, 2*, 48–51.

Pellegrini, A. D. (2001). Practitioner review: The role of direct observation in the assessment of young children. *Journal of Child Psychology and Psychiatry, 42*, 861–869.

Pellegrini, A. D. (2004). *Observing children in their natural worlds: A methodological primer*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Pellegrini, A. D., & Bartini, M. (2000). An empirical comparison of methods of sampling aggression and victimization in school settings. *Journal of Educational Psychology, 92*, 360–366.

Pellegrini, A. D., Ostrov, J. M., Roseth, C., Solberg, D., & Dupuis, D. (in press). Using observational methods to study children's and adolescents' development. In G. Melton, A. Ben-Arich, & J. Cashmore (Eds.). *Handbook of child research*. Beverly Hills, CA: Sage.

Pepler, D. J., & Craig, W. M. (1995). A peek behind the fence: Naturalistic observations of aggressive children with remote audiovisual recording. *Developmental Psychology, 31*, 548–553.

Pepler, D. J., Craig, W. M., & Roberts, W. L. (1998). Observations of aggressive and nonaggressive children on the school playground. *Merrill-Palmer Quarterly, 44*, 55–76.

Quake-Rapp, C., Miller, B., Ananthan, G., & Chiu, E-C. (2008). Direct observation as a means of assessing frequency of maladaptive behavior in youths with severe emotional and behavioral disorder. *The American Journal of Occupational Therapy, 62*, 206–211.

Sharpe, T. L. & Koperwas, J. (2003). *Behavior and sequential analyses: Principles and practice*. Thousand Oaks, CA: Sage Publications.

Shaw, D. S., Winslow, E. B., Owens, E. B., Vondra, J. I., Cohn, J. F., & Bell, R. Q. (1998). The development of early externalizing problems among children from low-income families: A transformational perspective. *Journal of Abnormal Child Psychology, 26*, 95–107.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sidener, T. M., Shabani, D. B., & Carr, J. E. (2004). A review of the Behavioral Evaluation Strategy and Taxonomy (BEST) Software Application. *Behavioral Interventions, 19*, 275–285.

Silk, J. B., Cheney, D. L., & Seyfarth, R. M. (1996). The form and function of post-conflict interactions between female baboons. *Animal Behaviour, 52*, 259–268.

Slee, P. T. (1987). *Child observation skills*. London, UK: Croom Helm.

Smith, G. A. (1986). Observer drift: A drifting definition. *The Behavior Analyst, 9*, 127–128.

Soukup, J. H., Wehmeyer, M. L., Bashinski, S. M., & Boyaird, J. A. (2007). Classroom variables and access to the general curriculum for students with disabilities. *Exceptional Children, 24*, 101–120.

Stangor, C. (2011). *Research methods for the behavioral sciences (4th ed)*. Belmont CA: Wadsworth.

Stauffacher, K., & DeHart, G. (2005). Preschoolers' relational aggression with siblings and friends. *Early Education and Development, 16*, 185–206.

Susser, S. A., & Keating, C. F., (1990). Adult sex role orientation and perceptions of aggressive interactions between boys and girls. *Sex Roles, 23*, 147–155.

Tapp, J.T., & Walden, T. (1993). PROCODER: A professional tape control, coding, and analysis system for behavioral research using videotape. *Behavior Research Methods, Instruments, & Computers, 25*. 53–56.

Tapp, J. T., Wehby, J. H., & Ellis, D. (1995). A Multi-option observation system for experimental studies: *MOOSES*. *Behavior Research Methods, Instruments, & Computers, 27*. 25–31.